Journal of the Korean Statistical Society (2001), 30: 2, pp 231–246

On Adaptation to Sparse Design in Bivariate Local Linear Regression

Peter Hall[1], Burkhardt Seifert[1,2] and Berwin A. Turlach[1,3]

ABSTRACT

Local linear smoothing enjoys several excellent theoretical and numerical properties, and in a range of applications is the method most frequently chosen for fitting curves to noisy data. Nevertheless, it suffers numerical problems in places where the distribution of design points (often called predictors, or explanatory variables) is sparse. In the case of univariate design, several remedies have been proposed for overcoming this problem, of which one involves adding additional "pseudo" design points in places where the original design points were too widely separated. This approach is particularly well suited to treating sparse bivariate design problems, and in fact attractive, elegant geometric analogues of univariate imputation and interpolation rules are appropriate for that case. In the present paper we introduce and develop pseudo data rules for bivariate design, and apply them to real data.

Keywords: Imputation; Interpolation; Kernel Methods; Nonparametric Regression; Pseudo Data; Smoothing; Tessellation; Triangulation.

1. INTRODUCTION

Problems of nonparametric regression with multivariate design points arise with increasing frequency in a range of applications, including dimension-reduction methods such as projection pursuit and ACE (e.g. Friedman and Stuetzle 1981, Breiman and Friedman 1985, Huber 1985), flexible multivariate models for high-dimensional data (e.g. Friedman 1988, 1991; Friedman and Silverman 1989), and generalized additive models (e.g. Hastie and Tibshirani 1986). In one dimension,

[1]Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

[2]Department of Biostatistics, University of Zurich, Zurich 8006, Switzerland

[3]Department of Mathematics and Statistics, The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia.

the virtues of local linear smoothing are well-known (e.g. Cleveland and Devlin 1988, Fan 1993, Hastie and Loader 1993). Multivariate generalisations of theoretical properties have been discussed by Ruppert and Wand (1994), in the context of local polynomial smoothing. However, both the attractive features (such as excellent theoretical performance) and the disadvantages (including difficulties with sparse design) of local linear smoothing generalise. Indeed, the "curse of dimensionality" correctly predicts that problems arising from sparse design increase with the number of dimensions.

In the univariate case a number of remedies have been proposed for dealing with the sparse design problem. They include ridging and bandwidth adjustment procedures, such as those discussed by Seifert and Gasser (1996), and local interpolation methods, considered for example by Hall and Turlach (1997). The latter are arguably the more promising in multivariate settings, not least because, on account of the "curse", bandwidth adjustment procedures for sparse data are more prone to error in two or more dimensions than they are with one. They are of long standing in statistics, since the "averaging" nature of interpolation methods assists in reducing the stochastic component of estimators. This virtue of interpolation has been exploited to analyse noisy data for well over a century, dating from at least the work of Cauchy 160 years ago. See also Steffensen (1950) and Tukey (1977).

In the present paper we propose geometric methods for determining when data are too sparse, for imputing "pseudo" design points when they are sparse, and for interpolating among the original data to determine the ordinates that should be associated with pseudo design points. We describe a number of different variants of the methods, outline their theoretical properties, and describe their performance for two real data sets.

Our algorithms are appropriate for any multivariate local polynomial regression of any degree, although we shall consider them in detail only for local linear smoothing and bivariate design. They are needed because general local polynomial regression involves division by a matrix (the "$X^T W X$" matrix, in the notation of weighted linear models) which is only positive semi-definite, not positive definite. (Even positive semi-definiteness requires that the kernel used to produce weights be nonnegative.) When the kernel is nonnegative the matrix's determinant can be guaranteed to be bounded away from zero by ensuring that a subset of the design points within a neighbourhood of the point at which the estimation takes place is of at least a certain size, and has at least a certain specified geometric configuration. In the case of local linear regression we shall show that

the minimum required number of design points is three, and that the necessary configuration is that they form a triangle whose area is bounded from below by a constant multiple of the square of bandwidth. Larger numbers of points, and more complex configurations, are appropriate in other circumstances.

Once the necessary number of points and their configuration have been determined, one should determine a rule for imputing new design points — that is, for adding pseudo design points in sufficient numbers, and in such a manner, that the minimal number-and-configuration rule is satisfied everywhere in the design space, even beyond the boundaries if that is necessary for overcoming potential boundary problems. There are several approaches to imputation. At one extreme one may conduct a painstaking search of the design space, to determine all places where, should one wish to estimate the regression mean there, problems will arise because the number-and-configuration rule is violated; and then determine a way of adding the bare minimum of extra design points so as to overcome the problem.

While this procedure might be regarded as statistically efficient, in that it minimises the number of pseudo data needed, it can be computationally expensive. A more practical rule is to divide the sample space into a regular lattice of tiles, and add a pseudodatum to each tile which does not already include a design point from the original design set. If the dimensions of the tiles are chosen appropriately, depending on the degree of the fitted local polynomial and on the bandwidth employed for the kernel weights, then one may ensure that the number-and-configuration rule holds regardless of what happens in other tiles. In the case of k–variate design the tiles are actually cells, for example the k–dimensional cubes in a lattice of such sets in $\mathbb{R}^k$.

The required dimensions of tiles depend on their shapes, in a curious way that demands non-standard geometric analysis to produce an optimal tiling scheme. In the case of local linear smoothing and bivariate design, the optimal tile shape is hexagonal. Here, optimality is defined in terms of maximising tile area; this ensures that the least expected number of pseudo data is required. Hexagonal tilings are a little more difficult than rectangular tilings to employ in practice, however. Since they are only marginally more efficient than rectangular tilings then we recommend any one of several easily-computed competitors of the latter type.

Following imputation, the next step is to determine the ordinates associated with pseudo design points, using an appropriate interpolation scheme. There are many interpolation algorithms, of which we consider only two, founded on Dirichlet tessellation and Delaunay triangulation, respectively.

The main purpose of this paper is to introduce imputation and interpolation rules in two dimensions, and describe their numerical performance. A more detailed theoretical study would show that they have all the hoped-for properties in connection with common methods for choosing bandwidth. For example, in the setting of standard asymptotic analysis (where sample size increases and other parameter settings remain fixed) they do not interact adversely with cross-validation or plug-in rules for bandwidth choice, no matter whether these are global or local in character. Bearing in mind that data sparseness problems usually only affect a small proportion of the region over which we aim to estimate a regression mean, we see that imputation and interpolation rules generally have negligible impact on a wide variety of global bandwidth choice methods.

It is sometimes proposed that problems of data sparseness, be they in one or higher dimensions, might be overcome by an adequate local bandwidth choice method, such as that based on near neighbours in the design data. There are, however, alternative viewpoints. Any local bandwidth selector that has good performance for estimating a regression mean should take into account information about that function, as well as about the design sequence. By way of contrast, nearest neighbour methods address only variation in the design sequence. Moreover, since local bandwidth choice techniques are based on only a relatively small fraction of the information in a sample, in particular that in the close vicinity of the point at which inference is being conducted, they are particularly susceptible to problems of data sparseness. As a result, they can exhibit very high variability in places where the design sequence is sparse. Using an imputation-and-interpolation rule, before making either a local or a global bandwidth selection, is one way of avoiding these problems.

Section 2 introduces our methods in the special case of local linear smoothing and for certain specific approaches to imputation and interpolation. Numerical performance is described in Section 3. Theoretical properties are noted briefly in Section 4, and technical details are deferred to an appendix.

2. METHODOLOGY

2.1. Model and Basic Estimator

In the bivariate case we observe independent and identically distributed triples $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, where the $\boldsymbol{X}_i$'s are bivariate predictors (the design points) and have a continuous distribution in $\mathbb{R}^2$ with density f, and the Y_i's are real–valued responses. The surface $g(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$ is assumed to be

smooth, and for the sake of simplicity of notation the residuals $Y_i - g(\boldsymbol{X}_i)$ are assumed to have constant variance σ^2 conditional on $\boldsymbol{X}_i$.

We shall use local linear methods, with kernel weights, to estimate $g(\boldsymbol{x})$, where $\boldsymbol{x}$ lies in the support of f. The kernel, K, will be taken to be a nonnegative function of Euclidean distance $\|\cdot\|$ on [0,1]. We assume that an appropriate linear transformation has been applied to the data prior to analysis, producing a bandwidth matrix $\boldsymbol{H} = h\,\boldsymbol{I}$ proportional to the identity. (This is done purely for notational simplicity. More generally, we could allow for unequal values in the diagonal, a "full" bandwidth matrix or locally varying bandwidths.) In this case, the local linear estimator $\hat{g}(\boldsymbol{x})$ of $g(\boldsymbol{x})$ becomes

$$\hat{g}(\boldsymbol{x}) = \boldsymbol{e}_1^T\,\widehat{\boldsymbol{\beta}}\,, \tag{2.1}$$

where $\widehat{\boldsymbol{\beta}}$ is the solution of

$$\boldsymbol{D}(\boldsymbol{x})\,\widehat{\boldsymbol{\beta}} = \boldsymbol{Z}(\boldsymbol{x}) \tag{2.2}$$

and, defining $K_i = K(\|\boldsymbol{X}_i - \boldsymbol{x}\| \,/\, h)$, the 3×3 matrix $\boldsymbol{D}(\boldsymbol{x})$ is defined by

$$\boldsymbol{D}(\boldsymbol{x}) = \begin{pmatrix} \sum_i K_i & \sum_i (\boldsymbol{X}_i - \boldsymbol{x})^T K_i \\ \sum_i (\boldsymbol{X}_i - \boldsymbol{x})\, K_i & \sum_i (\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^T K_i \end{pmatrix},$$

the 3×1 vector $\boldsymbol{Z}(\boldsymbol{x})$ is given by $\boldsymbol{Z}(\boldsymbol{x}) = \left(\sum_i Y_i K_i, \sum_i (\boldsymbol{X}_i - \boldsymbol{x})^T Y_i K_i\right)^T$, and $\boldsymbol{e}_1$ is the first 3×1 unit vector.

2.2. Imputation Rules

Let us assume temporarily that the original design sequence has been augmented by adding additional pseudo data, and that it is now $\{\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_{n^*}^*\} \supseteq \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, where $n^* - n \geq 0$ denotes the number of imputed design points. If the estimator $\hat{g}(\boldsymbol{x})$ were calculated for these design points, paired with ordinates Y_i^*, then the determinant $|\boldsymbol{D}(\boldsymbol{x})|$ would equal

$$\begin{aligned} |\boldsymbol{D}(\boldsymbol{x})| &= \sum_{1 \leq i < j < k \leq n^*} \left| \begin{pmatrix} \boldsymbol{X}_j^* - \boldsymbol{X}_i^*, & \boldsymbol{X}_k^* - \boldsymbol{X}_i^* \end{pmatrix} \right| K_i\, K_j\, K_k \\ &= \sum_{1 \leq i < j < k \leq n^*} \sin^2(\theta_{ijk})\, \|\boldsymbol{X}_j^* - \boldsymbol{X}_i^*\|^2\, \|\boldsymbol{X}_k^* - \boldsymbol{X}_i^*\|^2 K_i\, K_j\, K_k\,, \end{aligned}$$

where $(\cdot, \cdot)$ denotes inner product and θ_{ijk} represents the angle between $\boldsymbol{X}_j^* - \boldsymbol{X}_i^*$ and $\boldsymbol{X}_k^* - \boldsymbol{X}_i^*$. This formula can be deduced from a result on determinants of the product of rectangular matrices (e.g. Rao, 1976, p. 33). Inspecting this formula

we see that if K is bounded away from zero on $[0, 1-\delta]$ for each $\delta > 0$ then a sufficient condition for $|\boldsymbol{D}(\boldsymbol{x})| > C\,h^4$ (for a constant C depending on K) is that for some $0 < r < 1$ there exists $A = A(r)$ such that

> any circle of radius rh contains at least three design points, pseudo or original, which form a triangle whose area is bounded below by Ah^2. (2.3)

A set of data and pseudo data which satisfies this condition will produce a matrix $D(\boldsymbol{x})$ that is guaranteed to be invertible, and hence such that equation (2.2) is well conditioned. This alleviates many of the problems of sparse design.

To illustrate possibilities we consider three imputation algorithms that ensure condition (2.3). The first is preferable from a computational viewpoint, we argue, and so is given in more detail than the other two.

Rule 1: Tiling method. Divide the plane into a regular pattern of "tiles", represented by a tile type $\mathcal{T}$ (e.g. hexagonal, square or triangular) and tile area C_1h^2; and add a pseudo design point at the centre of any tile which does not already contain any design points. We may select $C_2 = C_2(\mathcal{T}) > 0$ such that whenever $C_1 < C_2$, any circle of radius $rh < h$ placed into the plane contains at least three design points (real data or pseudo data) forming a triangle whose area is not less than Ah^2, where $A = A(\mathcal{T}, r) > 0$. An advantage of this algorithm over that considered next is that it requires only a systematic search, over a completely determined and finite number of possible tiles, rather than an indeterminate sequential search over a potential continuum of candidates for $\boldsymbol{x}$.

Figure 2.1 illustrates tilings of different types $\mathcal{T}$, and Table 2.1 lists the quantities (tile edge width, or tile area) that determine the supremum of possible values of $C_2(\mathcal{T})$ — or equivalently, the supremum of tile dimensions that permit (2.3) to be satisfied for any $0 < r < 1$.

Table 2.1: Maximal sidelengths and areas $C_2(\mathcal{T})$ for different tilings that guarantee three points in each circle of radius $rh < h$, defining a triangle of area bounded above const. h^2.

Tiling type, $\mathcal{T}$	Edge length / h	Area / h^2
Regular square	$1/\sqrt{5}$	0.200
Bricked square	$32/\sqrt{4745}$	0.216
Bricked $(\sqrt{3} : 2)$–rectangle	$18/\sqrt{1281}$	0.219
Hexagonal	13/43	0.237

In principle, a choice of tiling that maximizes face area is preferable, since it is associated with least average requirement for generating pseudo data. In this respect, triangular tilings may be shown to be inferior to square ones. Additionally they are relatively awkward to implement, and so are not represented in the figure or table. Those tilings that are represented are listed in order of increasing face area: regular square, bricked square, bricked rectangle and hexagon. The latter is not computationally attractive, however, and is not recommended for this reason.

For the bricked rectangular tiling, the longer side length is listed in the table. We did not attempt to optimise bricked rectangular tilings over all ratios of side lengths. The chosen side ratio produces rectangles whose centres are on a grid of equilateral triangles, and appears to be close to the optimum.

Clearly, the location and orientation of the tiling have an influence on the placing of pseudo design points. Our experience indicates that this is negligible, however.

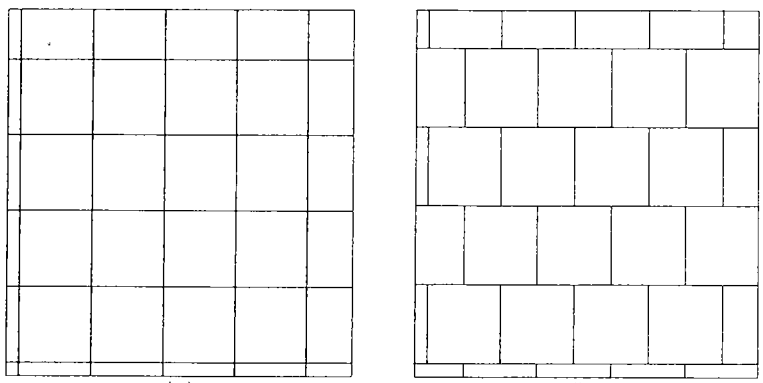


Figure 2.1: Examples of regular tilings: (a) regular square, (b) bricked square.

Rule 2: Continuum search method. Here we select an appropriate number $r_1 > 0$, and conduct a sequential numerical search in $\mathbb{R}^2$ for all points $\boldsymbol{x}$ such that the circle $\mathcal{B}(\boldsymbol{x})$ of radius r_1 centred at $\boldsymbol{x}$ contains no design points, adding a pseudo design point at $\boldsymbol{x}$ if in fact there exists no $\boldsymbol{X}_i$ in $\mathcal{B}(\boldsymbol{x})$, before passing to the next candidate for $\boldsymbol{x}$. This rule produces a set of data and pseudo data satisfying condition (2.3), but is not computationally attractive because of the high level of labour involved.

It is sufficient to take $r_1 = r/3 < 1/3$, where r is as in (2.3). This is made clear by the following theorem, which is useful in deriving a range of different

imputation rules. Its proof is based on a simple geometric argument, and is not given here.

Theorem 2.1. *Let $r_1 < r/2$, and suppose every circle of radius $r_1 h$ contains at least one design point. Then, every circle of radius $r h$ contains at least three design points, which form a triangle with area bounded from below by $(r-2r_1)^2 h^2$. Moreover, the centre $\boldsymbol{x}$ is inside the triangle formed by these points.*

Rule 3: Edge interpolation method. Construct the Delaunay triangulation of the convex hull of the set of design points. Divide each edge of the triangulation into N equal parts, where N equals the smallest integer not less than $l/(rh)$, $r < 1$, and l equals the edge length. Place a pseudo design point at each of the points of division. Once this operation has been completed for all edges, recompute the Delaunay triangulation for the new set of design and pseudo design points, and repeat the division step. Continue iterating until a triangulation is produced that has no edges longer than rh. At the termination of this algorithm, condition (2.3) is satisfied on the set of all points that are within the convex hull of the original design points or at most rh away of it. The condition will be satisfied over any region of interest, not just within the convex hull of the design points, if one uses a triangulation such as that offered by the triangulation program "Deldir" of Turner and McQueen (1996).

2.3. Interpolation rules

There is a variety of ways of using regular data triples $(\boldsymbol{X}_i, Y_i)$ so as to derive the value Y_j^* associated with a pseudo design point $\boldsymbol{X}_j^*$. We shall consider two, founded on the Dirichlet tessellation and Delaunay triangulation respectively. See Ripley (1981, pp. 38 ff) for an account of geometric properties of these schemes.
Rule 1: Dirichlet tessellation. Here we take $Y_j^* = Y_i$, where the index i is chosen so that $\boldsymbol{X}_i$ is the germ or centre of the unique Dirichlet cell (or tile) in which $\boldsymbol{X}_j^*$ lies. We take corresponding mean values of observations in adjacent cells if $\boldsymbol{X}_j^*$ lies on a cell boundary. The Dirichlet tessellation rule automatically produces a form of horizontal extrapolation outside the convex hull of the original design sequence.
Rule 2: Delaunay triangulation. Here, Y_j^* is the linear interpolant of those values Y_i paired with the three $\boldsymbol{X}_i$'s at the vertices of the Delaunay triangle within which $\boldsymbol{X}_j^*$ lies, with the obvious generalization on edges. (If one is using the third data imputation rule introduced above then the interpolant may be derived internally

to the algorithm.) Outside the convex hull we suggest using a form of horizontal extrapolation, to minimize the problems of noise. One approach is to employ Y_j–values obtained from the Dirichlet tessellation rule; another is to project $\boldsymbol{X}_j^*$ onto the convex hull of the original design sequence, and use the weighted mean of the two adjacent observations (if the projection lies on an edge), and the single observation (if the projection is a single design point — this is not an event of measure zero).

In practice we noticed that the interpolation rule based on the Dirichlet tessellation produces, as one may expect, values Y_j^* at the pseudo design points $\boldsymbol{X}_j^*$ which are highly discontinuous along the borders of the Dirichlet tiles. We found that this could have an adverse effect on the subsequent smoothing step, and on those grounds argue that the interpolation rule based on the Delaunay triangulation is preferable.

Likewise, for the case of Delaunay triangulation, imputing values Y_j^* at pseudo design points $\boldsymbol{X}_j^*$ lying outside the convex hull of the original data using Y_j^*'s given by the Dirichlet tessellation rule produces erratic values Y_j^* along the convex hull, adversely affecting the final smooth. Hence, we recommend the second approach, which projects each $\boldsymbol{X}_j^*$ onto the convex hull and takes as Y_j^* the value interpolated at this point.

Since the imputed data are not independent of one another then the usual approaches to estimating the variance of an estimator of a regression mean will not be valid. However, we argue that the "usual" approaches are generally not valid in the context of sparse design, since they rely on asymptotic arguments that do not admit sparsity.

3. NUMERICAL EVALUATION

To implement the methods discussed above we adapted the triangulation program "Deldir" of Turner and McQueen (1996), which generates tessellations using an algorithm of Lee and Schacter (1980). See also Green and Sibson (1978).

When using Delaunay triangulation, choice of the tiling rule (see Section 2.2) has only a minor effect on the final smooth. On grounds of computational simplicity we prefer the regular square tiling illustrated in panel (a) of Figure 2.1. We shall illustrate the application of this choice of imputation and interpolation rules to two data sets.

The data for our first example come from the Clinic of Neonatology in the University Hospital, Zürich. Birth weight and gestational age (length of preg-

nancy) were measured on 29 babies, to predict cerebral blood volume. Since the scales of the two predictor variable are quite different, we rescaled both so that they had empirical variance 1. Next, we imputed pseudo design points $\boldsymbol{X}_j^*$ using a regular square tiling, and computed the associated values Y_j^* using the Delaunay triangulation rule. Finally, we employed a bivariate local linear smoother with bandwidth $h = 0.5$ to fit a surface to the data. Figure 3.1 shows the contours of the fitted surface, with the predictor variables transformed back to their original scale.

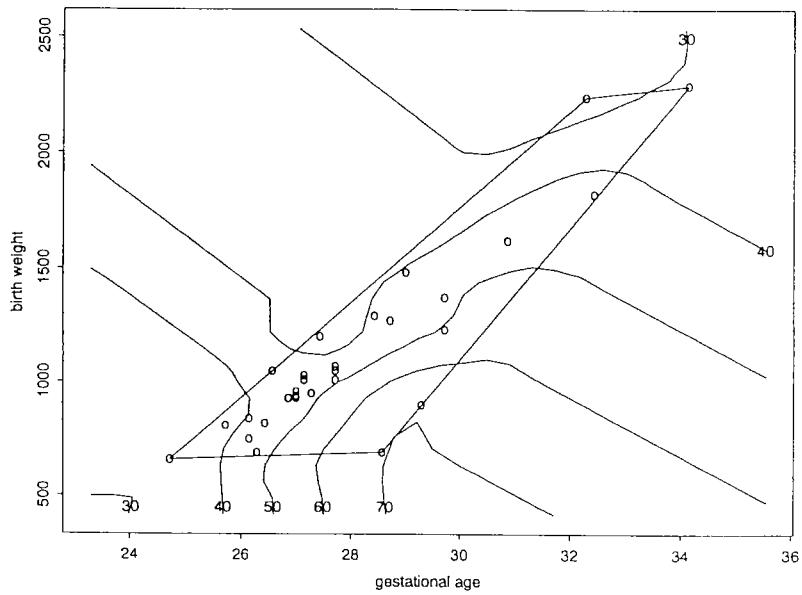


Figure 3.1: Contour lines for the surface fitted to the neonatal data by local linear smoothing. Data points are denoted by circles, and the polygon depicts the convex hull.

In this example (as in the next one) the bandwidth was chosen visually (in ongoing work we are studying the adaptation of bandwidth selectors for this method). Varying the bandwidth does not change the situation greatly. Increasing the bandwidth pulls the contour lines further apart, since the surface becomes smoother. Decreasing the bandwidth pushes them together. In the latter case the lines also become more ragged and (at extremely small bandwidths) further local extrema appear which are clearly due to undersmoothing. But in both cases the general direction in which the contour lines are running remains the same. The number of imputed pseudo data had no significant influence on these results.

The second data set originated in a spatial problem discussed by O'Conner and Leach (1979). We took the data from Green and Silverman (1994, Chapter 7) and compared our method with the thin plate spline used by them. The data consist of 38 sampling points, collected from a mine in Cobar, Australia. The predictor variables $\boldsymbol{X}_j$ represent geographic locations where the "true widths", Y_j, of the ore–bearing layer were measured.

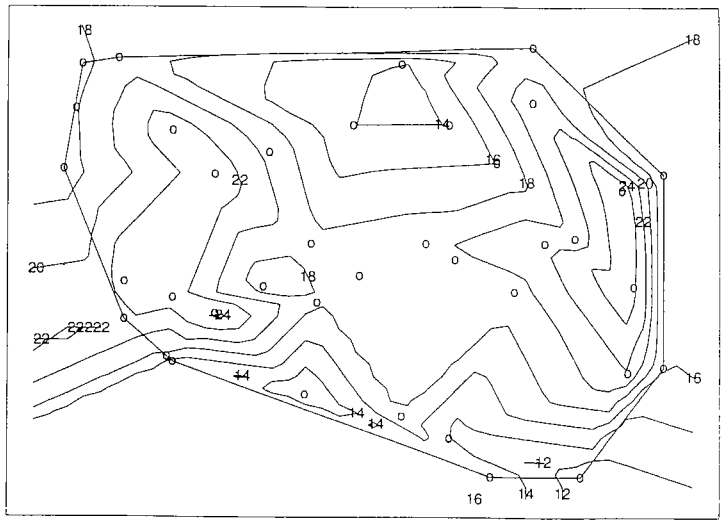
We used the procedure described for the neonatal data. Panel (a) of Figure 3.2 shows the contour lines calculated from the surface derived by interpolating the data. To obtain panel (b) the interpolating surface was smoothed using the bandwidth $h = 0.35$. Comparing these pictures with their counterparts in Green and Silverman (1994) we see that our method and theirs give similar results. The only major difference is in the mode located at the uppermost boundary of the pictures. Green and Silverman's thin-plate spline suggests that the contour lines at this mode have approximately a vertical axis, whereas the axis for the contours in panel (b) of Figure 3 is inclined to the vertical.

4. THEORETICAL PROPERTIES

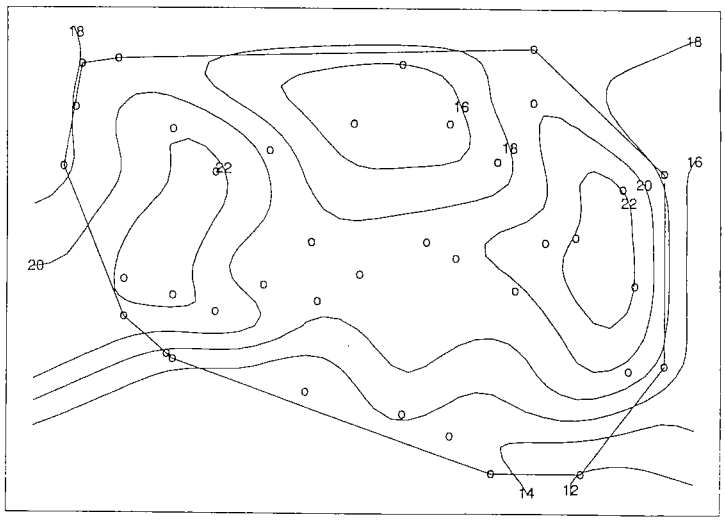
If the imputation or interpolation rules suggested in Section 2 are employed then all the classical results on mean–square properties of local linear regression in two dimensions are valid in an unconditional sense. In particular, asymptotic expansions of bias and variance (see e.g. Theorems 2.1 and 2.2 of Ruppert and Wand, 1994) are valid without need for conditioning on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, and with the subscript p removed from remainder terms $o_p(\ldots)$. No changes are required to the regularity conditions imposed by Ruppert and Wand on the design density and regression mean. The only condition needed on h is that for some $\epsilon > 0$ and all sufficiently large n, $n^{-(1/2)+\epsilon} \leq h \leq n^{-\epsilon}$. The only conditions required on K are those imposed in Section 2.2.

We may deduce from the results in the previous paragraph that the bias and variance of $\hat{g}$ are respectively of size h^2 and $(nh^2)^{-1}$, and that this holds uniformly in the support of the design density f, provided the support may be represented as a finite union of convex sets and that f is bounded away from zero on its support. Thus, our imputation-and-interpolation approach effectively removes all first-order problems associated with sparse design.

These properties are available because, under condition (2.3), the maximal possible variance of $\hat{g}(\boldsymbol{x})$ is bounded by a quantity that does not depend on the specific design configuration. Our next result makes this clear.



(a)



(b)

Figure 3.2: (a) Contour lines derived from the interpolated mine data. (b) Same as for panel (a), except that smoothing was applied using bandwidth $h = 0.35$. Data points are denoted by circles, and the outer polygon depicts the convex hull.

As in Section 2.2 we ask that the kernel K be bounded away from zero on $[0, 1-\delta]$ for each $\delta > 0$. Assume that the imputation rule satisfies (2.3), and adds no more than B, say, extra points to any circle of radius of h, where B depends only on the rule. (This condition is satisfied by each of rules produced by Rules 1–3 of Section 2.2.) Furthermore, suppose that the interpolation rule produces values Y_i^* that may be represented as $Y_i^* = \boldsymbol{a}_i^T \boldsymbol{Y}$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{a}_i$ is an n–vector satisfying $\boldsymbol{a}_i^T \boldsymbol{a}_i \leq 1$. (The interpolation rules suggested in Section 2.3 satisfy this requirement.)

Theorem 4.1. *Under the above conditions there exists a constant C, depending only on A (from (2.3)), B, r and the kernel K, such that*

$$\text{var}\,\{\hat{g}(\boldsymbol{x})|\boldsymbol{X}\} \leq \sigma^2\, C\,.$$

If the three points mentioned in condition (2.3) may always be chosen to be vertices of a triangle that contains $\boldsymbol{x}$, then it suffices to choose

$$C = \sigma^2\,(B+1)\,\frac{\max\limits_{u\in[0,1]} K(u)}{\min\limits_{u\in[0,r]} K(u)}.$$

Related results have been derived by Mammen and Marron (1996) in the case of the shifted Nadaraya–Watson estimator, and by Herrmann (1996) for a modified Gasser–Müller estimator. Any one of the imputation rules presented in Section 2 may be constructed such that they fulfill the second part of the Theorem. (In the case of the rule determined by Rule 1, for example, it suffices to choose tile width to be a sufficiently small multiple of h. Rule 2 already guarantees that $\boldsymbol{x}$ is in the triangle; see Theorem 2.1.)

Appendix A: Proof of Theorem 4.1

The conditional variance of $\hat{g}(\boldsymbol{x})$ is

$$\begin{aligned}
\text{var}\,\{\hat{g}\{\boldsymbol{x}\}\} &= \text{var}\,(\boldsymbol{e}_1^T\, \boldsymbol{D}(\boldsymbol{x})^{-1}\, \boldsymbol{Z}(\boldsymbol{x})) \\
&= \boldsymbol{e}_1^T \left(\boldsymbol{X}^{*T}\, \boldsymbol{W}^*\, \boldsymbol{X}^*\right)^{-1} \left\{\boldsymbol{X}^{*T}\, \boldsymbol{W}^*\, \text{Cov}(\boldsymbol{Y}^*)\, \boldsymbol{W}^*\, \boldsymbol{X}^*\right\} \left(\boldsymbol{X}^{*T}\, \boldsymbol{W}^*\, \boldsymbol{X}^*\right)^{-1} \boldsymbol{e}_1 \\
&\leq \lambda_{\max}\{\text{Cov}(\boldsymbol{Y}^*)\}\, \lambda_{\max}(\boldsymbol{W}^*)\, \boldsymbol{e}_1^T \left(\boldsymbol{X}^{*T}\, \boldsymbol{W}^*\, \boldsymbol{X}^*\right)^{-1} \boldsymbol{e}_1\,,
\end{aligned}$$

where $\lambda_{\max}(\cdot)$ denote the maximum eigenvalue operator, and $\boldsymbol{W}^*$ is the matrix of weights computed from the original and imputed design points. Now,

$$\lambda_{\max}(\boldsymbol{W}^*) \leq \max_{u\in[0,1]} K(u)\,,$$

and, since pseudo observations are of the form $Y_i^* = \boldsymbol{a}_i^T \boldsymbol{Y}$ with $\boldsymbol{a}_i^T \boldsymbol{a}_i \leq 1$,

$$\lambda_{\max}\{\mathrm{Cov}(\boldsymbol{Y}^*)\} \leq \sigma^2 (B+1) .$$

From the Rao–Blackwell theorem, $\boldsymbol{e}_1^T (\boldsymbol{X}^{*T} \boldsymbol{W}^* \boldsymbol{X}^*)^{-1} \boldsymbol{e}_1$ would be the minimal variance of unbiased linear estimators for $\boldsymbol{e}_1^T \boldsymbol{\beta}$ given $\mathrm{Cov}(\boldsymbol{Y}^*) = (\boldsymbol{W}^*)^{-1}$. One of the estimators under comparison, $\tilde{g}$ say, is the linear interpolant of three points $\boldsymbol{X}_i^*, \boldsymbol{X}_j^*, \boldsymbol{X}_k^*$, say, mentioned in condition (2.3). Since the area of the triangle is bounded from below by a constant multiple of h^2, so also the sides are bounded from below by a constant multiple of h. Hence, $\tilde{g}(\boldsymbol{X}_i^*, \boldsymbol{X}_j^*, \boldsymbol{X}_k^*) = \alpha_i Y_i^* + \alpha_j Y_j^* + \alpha_k Y_k^*$, with the weights α bounded above by a constant depending only on A, r and K. Therefore, $\boldsymbol{e}_1^T (\boldsymbol{X}^{*T} \boldsymbol{W}^* \boldsymbol{X}^*)^{-1} \boldsymbol{e}_1$ is bounded above by a constant depending only on these quantities. The theorem follows from these results.

If $\boldsymbol{x}$ is inside the triangle, the weights α are nonnegative and sum to one, whence

$$\boldsymbol{e}_1^T (\boldsymbol{X}^{*T} \boldsymbol{W}^* \boldsymbol{X}^*)^{-1} \boldsymbol{e}_1 \leq \frac{\alpha_i^2}{K_i} + \frac{\alpha_j^2}{K_j} + \frac{\alpha_k^2}{K_k} \leq \frac{1}{\min\limits_{u \in [0,r]} K(u)} ,$$

which proves the second part of the theorem.

References

Breiman, L. and Friedman, J.H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation (with discussion)," *Journal of the American Statistical Association*, 80, 580–619.

Chu, C.-K. and Marron, J.S. (1991), "Choosing a Kernel Regression Estimator," *Statistical Science*, 6, 404–436.

Cleveland, W. and Devlin, S. (1988), "Locally Weighted Regression: an Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, 83, 596–610.

Fan, J. (1993), "Local Linear Regression Smoothers and their Minimax Efficiencies," *Annals of Statistics*, 21, 196–216.

Friedman, J.H. (1988), "Fitting Functions to Noisy Data in High Dimensions," in *Computing Science and Statistics: Proceedings of the Twentieth Symposium on the Interface*, eds E.J. Wegman, D.T. Gantz and J.J. Miller, 13–43. ASA, Alexandria, Virginia.

Friedman, J.H. (1991), "Multivariate Adaptive Regression Splines (with discussion)," *Annals of Statistics*, 19, 1–141.

Friedman, J.H. and Silverman, B.W. (1989), "Flexible Parsimonious Smoothing and Additive Modeling", *Technometrics*, 31, 3–39.

Friedman, J.H. and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

Green, P.J. and Sibson, R. (1978), "Computing Dirichlet Tessellations in the Plane," *Computer Journal*, 21, 168–173.

Green, P. J. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London.

Hall, P. and Turlach, B. (1997), "Interpolation methods for adapting to sparse design in nonparametric regression," *Journal of the American Statistical Association*, 92, 466-476.

Hastie, T. and Loader, C. (1993), "Local Regression: Automatic Kernel Carpentry," *Statistical Science*, 8, 120–143.

Hastie, T. and Tibshirani, R. (1986), "Generalized Additive Models (with discussion)," *Statistical Science* 1, 297–318.

Herrmann (1996), "On the convolution type kernel regression estimator", Preprint-Nr. 1833, Fachbereich MathematiK, TH Darmstadt, Germany.

Huber, P.J. (1985), "Projection Pursuit (with discussion)," *Annals of Statistics*, 13, 435–475.

Lee, D.T. and Schacter, B.J. (1980), "Two algorithms for constructing a Delaunay triangulation," *International Journal of Computer and Information Sciences*, 9, 219–242.

Mammen, E. and and Marron, J.S. (1996), "Mass Recentered Kernel Smoothers," Research Report No. 75, Sonderforschungsbereich 373, Humboldt–Universität zu Berlin.

O'Conner, D.P.H. and Leach, G.B. (1979), "Geostatistical analysis of 18CC Stope block, CSA mine, Cobar, NSW," *Estimation and Statement of Mineral Reserves*, Australian IMM, Melbourne, pp. 145–153.

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, Second Edition, Wiley, New York.

Ripley, B.D. (1981), *Spatial Statistics*, Wiley, New York.

Ruppert, D., and Wand, M.P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370.

Seifert, B. and Gasser, T. (1996), "Finite Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association*, 91, 267–275.

Steffensen, J.F. (1950). *Interpolation.* Chelsea, New York.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass.

Turner, R. and MacQueen, D. (1996), "S function Deldir to Compute the Dirichlet (Voronoi) Tesselation and Delaunay Triangulation of a Planar Set of Data Points." Available from `Statlib`.